To link to this article : DOI :10.1007/s11192-012-0862-y
URL : http://dx.doi.org/10.1007/s11192-012-0862-y

Experimenting with the partnership ability φ-index on a million computer scientists

Guillaume Cabanac

Abstract Schubert introduced the partnership ability φ-index relying on a researcher's number of co-authors and collaboration rate. As a Hirsch-type index, φ was expected to be consistent with Schubert–Glänzel's model of h-index. Schubert demonstrated this relationship with the 34 awardees of the Hevesy medal in the field of nuclear and radiochemistry ($r^2 = 0.8484$). In this paper, we upscale this study by testing the φ-index on a million researchers in computer science. We found that the Schubert–Glänzel's model correlates with the million empirical φ values ($r^2 = 0.8695$). In addition, machine learning through symbolic regression produces models whose accuracy does not exceed a 6.1 % gain ($r^2 = 0.9227$). These results suggest that the Schubert–Glänzel's model of φ-index is accurate and robust on the domain-wide bibliographic dataset of computer science.

Keywords Partnership ability index · Co-authorship · Empirical validation · Symbolic regression

Introduction

The literature of scientometrics features a wealth of indicators devoted to the measurement of individual performance (Bar-Ilan 2008). As a prominent author-based indicator, the h-index intends to measure the impact of an author's research according to his/her number of publications and citation rate (Hirsch 2005). Many variants have subsequently stemmed from the h-index (Alonso et al. 2009; Schreiber et al 2012). This article deals with one of these Hirsch-type indexes: the partnership ability φ-index devised by Schubert (2012a) with the principles of the h-index in mind. Rousseau (2012) also stressed its relation to the h-degree of nodes in weighted networks introduced in (Zhao et al 2011). The φ-index accounts for a researcher's number of co-authors and collaboration rate. As Schubert (2012a, p. 304) put it:

G. Cabanac (✉)
Computer Science Department, IRIT UMR 5505 CNRS, University of Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France
e-mail: guillaume.cabanac@univ-tlse3.fr

An author is said to have a co-author partnership ability φ, if with φ of his/her n co-authors [he/she] had at least φ joint papers each, and with the other $(n-\varphi)$ co-authors [he/she] had no more than φ joint papers each.

Schubert (2012a, b) also stressed the analogy between the basic properties of the φ-index and those of its prototype, Hirsch's (2005) *h*-index:

- $\varphi = 0$ if and only if the author had only single-authored papers.
- $\varphi = 1$ in one of the following cases:
  - (a) If the author had an arbitrary number of double-authored papers with the very same co-author each.
  - (b) If the author had an arbitrary number of co-authored papers with no co-authors occurring more than once.
  - (c) If the author had an arbitrary number of double-authored papers with the very same co-author each AND an arbitrary number of co-authored papers with no co-authors occurring more than once (Rousseau 2012).
- $\varphi > 1$ in all other cases.

Let us illustrate the φ-index with the case of Albert Einstein, who is credited with 272 journal articles. Only 44 of these (i.e., 16 %) were co-authored with colleagues. Table 1 shows his 24 co-authors with the number of joint papers per co-author. As represented by the dashed line Einstein has $\varphi = 3$, since with three of his co-authors he had at least three joint papers each, and with the other 21 co-authors he had no more than three joint papers each. Notice that Schubert (2012a, b) considers φ as a "natural" delimitation of closest co-authors, with the top section of the co-authors list (where rank $\leq \varphi$) named the "φ-core" of co-authors.

As a Hirsch-type index, φ was expected to be consistent with Glänzel's (2006) model of *h*-index, which had been further investigated in (Schubert and Glänzel 2007). Schubert (2012a) transposed Glänzel's (2006) model to the case of the partnership φ-index. The

Table 1 Co-authors of A. Einstein, with their number of co-authored journal papers and partnership rank

Rank	Name	Nbr of joint papers	Rank	Name	Nbr of joint papers
1	W. Mayer	8	13	P. Bergmann	1
2	W. J. de Haas	4	14	B. Cohen	1
3	N. Rosen	4	15	T. de Donder	1
4	L. Infeld	3	16	A. D. Fokker	1
5	J. Laub	3	17	M. Grossman	1
6	P. Ehrenfest	2	18	B. Hoffman	1
7	J. Grommer	2	19	H. Mühsam	1
8	L. Hopf	2	20	W. Pauli	1
9	B. Kaufman	2	21	W. Ritz	1
10	B. Podolosky	2	22	W. de Sitter	1
11	E. G. Straus	2	23	O. Stern	1
12	V. Bargmann	1	24	R. C. Tolman	1

http://en.wikipedia.org/wiki/List_of_scientific_publications_by_Albert_Einstein resulting φ^*_{SG} function (Eq. 1) is deemed to approximate the φ value of an author, based on three parameters: c is a positive constant of order one, a is the total number of co-authors, and z is the mean number of occurrence of the co-authors.

$$\varphi^*_{SG} = c \cdot a^{\frac{1}{3}} \cdot z^{\frac{2}{3}} \tag{1}$$

Regarding Einstein's collaborations shown in Table 1, one finds $a = 24$ distinct co-authors and $z = 1.9583$ co-authored papers per collaborator on average. Consequently, the φ-index of Einstein is evaluated by (Eq. 1) as $\varphi^*_{SG} = 1 \cdot 24^{\frac{1}{3}} \cdot 1.9583^{\frac{2}{3}} \approx 4.51$. In the case of Einstein, the approximation $\varphi^*_{\mathrm{SG}} = 4.51$ overestimates the empirical value of $\varphi = 3$. According to the model, Einstein was expected to have a greater partnership ability than he actually had.

In order to check the accuracy of the φ^*_{SG} model, Schubert (2012a) correlated the φ and φ^*_{SG} values computed for the 34 awardees of the Hevesy medal (1975–2011) in the field of nuclear and radiochemistry. He reported that φ^*_{SG} is consistent with φ on this dataset ($r^2 = 0.8484$), while stressing the need to confirm these results with larger bibliographic datasets from various fields and subfields of science. A subsequent study of partnership ability among 58 jazz performers (Schubert 2012b) also showed a strong support for the validity of the φ^*_{SG} model ($r^2 = 0.8845$).

Further to Schubert's (2012a) work, we tackle the following question in the present article: Is the φ^*_{SG} model still accurate for (1) a much larger sample of researchers characterized by (2) a larger range of expertise and (3) who were drawn from a whole field of science?

The article is organized as follows. We first introduce a publicly available dataset that records the bibliographies of more than one million computer scientists. Second, we correlate φ and φ^*_{SG} to evaluate the accuracy of φ^*_{SG}. Finally, we use symbolic regression to revise the parameters of φ^*_{SG} by learning from the considered dataset.

Data: bibliographical records of a million computer scientists

The Data bibliography and library project (DBLP) collects metadata about the scholarly publications in computer science (Ley 2002), starting from 1936. These are freely released as an XML file[1] of 1 GB in size. At the time we started the present study (12 March 2012), the DBLP was indexing 1,919,594 documents authored by 1,095,174 researchers. Several types of documents are recorded, such as books, PhD dissertations, journal articles, conference proceedings, and conference papers. The interested reader is referred to (Cabanac 2011) for a UML model of the metadata recorded by the DBLP.

For the present study, we focused on the two categories of referred papers that are acknowledged in computer science: papers published in journals or in the proceedings of workshops and conferences (Chen and Konstan 2010; Freyne et al 2010). These represent 1,833,417 papers authored by 1,072,213 researchers, who have a large range of expertise—from beginners to appraised experts in the field. Notice that the output of researchers in the DBLP fits Lotka's (1926) law, as previously shown in (Elmacioglu and Lee 2005). In the remainder of the paper, we refer to this dataset as `DBLP_2012`.

Among the computer scientists recorded in the DBLP, Cabanac (2012) identified those 2,849 researchers who serve as gatekeepers for the 77 core journals in Information Systems, which is a subfield of computer science. In the remainder of the paper, we refer to this dataset as `EB_IS_2009`.

[1] http://dblp.uni-trier.de/xml

Assessing the accuracy of Schubert–Glänzel's φ^*_{SG} model of the φ-index

The values of the empirical (φ) and theoretical (φ^*_{SG}) partnership ability index were computed for the 1,072,213 researchers (see Appendix). Then, we correlated the φ and φ^*_{SG} variables. The coefficient of determination $r^2 \in [0, 1]$ was used to measure the accuracy ($r^2 \rightarrow 1$) of Schubert–Glänzel's model with respect to empirical data. In this section, we report the results obtained with the two aforementioned datasets: `EB_IS_2009` and `DBLP_2012`.

Testing Schubert–Glänzel's φ^*_{SG} model with a sample of 2,849 gatekeepers

Schubert (2012a) showed that φ^*_{SG} leads to a good approximation ($r^2 = 0.8484$) of φ for the 34 awardees of the Hevesy medal. We upscaled this experiment with the 2,849 gatekeepers of the `EB_IS_2009` dataset. These are acknowledged researchers, thus with quite similar profiles to those of the Hevesy medal awardees.

The linear regression between the two variables is shown in Fig. 1. The coefficient of determination $r^2 = 0.9211$ shows a very strong relation between the two variables. This is a confirmation of Schubert's (2012a) results: φ^*_{SG} produces a good approximation of φ for leading researchers. The accuracy of the approximation is even 8.6 % better on the `EB_IS_2009` dataset compared to the Hevesy dataset.

Testing Schubert–Glänzel's φ^*_{SG} model with a million computer scientists

In this section we measure the accuracy of the φ^*_{SG} model applied to `DBLP_2012`, as a more diverse and 376-fold larger sample of researchers than `EB_IS_ 2009`. Figure 2 shows the linear regression between φ^*_{SG} and φ on the `DBLP_2012` dataset of 1,072,213 computer scientists. The $r^2 = 0.8695$ value is 2.5 % higher than the r^2 value reported in

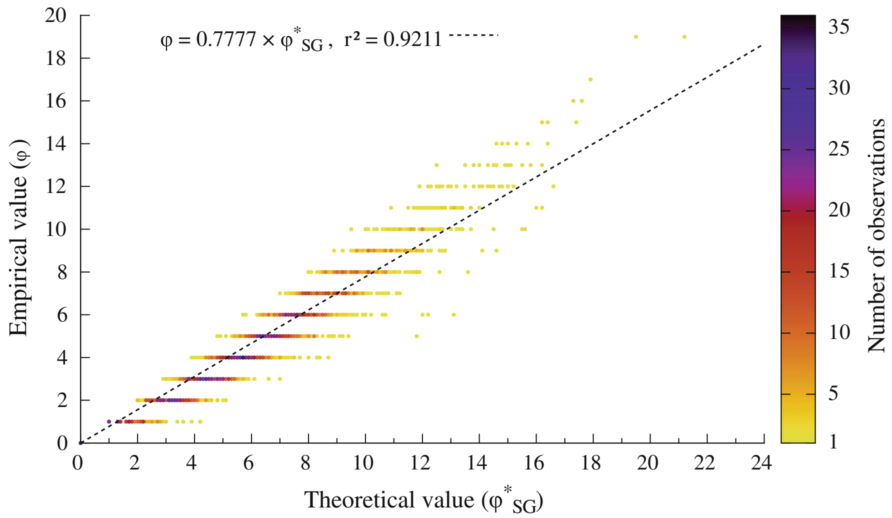

Fig. 1 Linear fit between the theoretical (φ^*_{SG}) and empirical (φ) values of the partnership ability index for the 2,849 gatekeepers of the `EB_IS_2009` dataset. Points are *colored* according to their density: *light points* show fewer observations than *dark points*, which show a larger number of observations

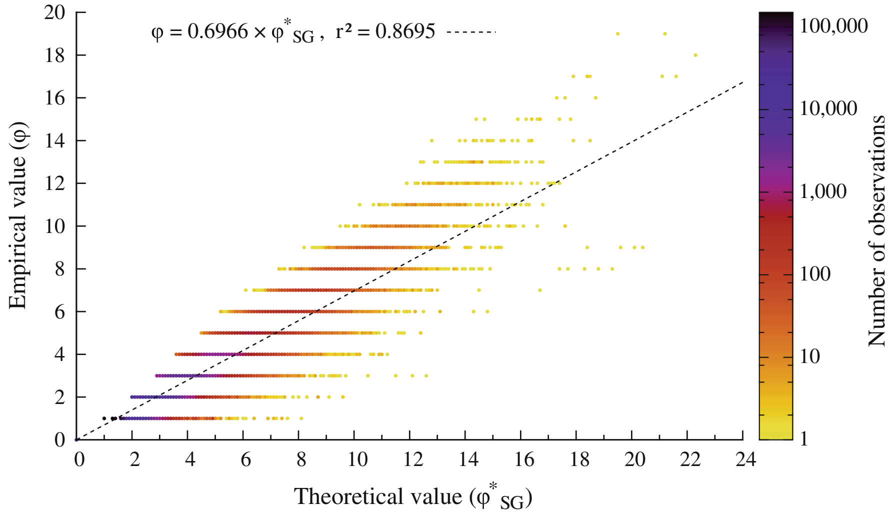

Fig. 2 Linear fit between the theoretical (φ^*_{SG}) and empirical (φ) values of the partnership ability index for the 1,072,213 authors of the `DBLP_2012` dataset. Points are *colored* according to their density: *light points* show fewer observations than *dark points*, which show a larger number of observations

(Schubert 2012a). This result suggests that φ^*_{SG} is a good approximation of φ for leading researchers and less prominent, mainstream researchers alike.

Revising the parameters of Schubert–Glänzel's φ^*_{SG} model through machine learning

Given the two datasets, we intended to check whether we could revise the parameters of φ^*_{SG} to increase its accuracy. We relied on symbolic regression (Koza 1992), as a machine learning approach used to discover models (i.e., formulas) from input data. This approach is inspired by biological evolution. Several random formulas involving operands (e.g., a and z) and operators (e.g., multiplication, square root) are first generated. Then, the best solutions according to a fitness function (e.g., r^2) are selected. Finally, a new generation of formulas are generated by combining the former ones. This process is repeated until a user-defined fitness threshold is reached.

We used symbolic regression to optimize the parameters of (Eq. 1) with respect to the coefficient of determination r^2. In this section, we report the results of the Eureqa[2] software that implements symbolic regression (Schmidt and Lipson 2009). The parameters of φ^*_{G} (Eq. 2) were learned on the `EB_IS_2009` dataset, while the parameters of φ^*_{D} (Eq. 3) were learned on the `DBLP_2012` dataset.

$$\varphi^*_G = 0.5248 \cdot a^{0.3982} \cdot z^{0.7743} \tag{2}$$

$$\varphi^*_D = 0.6546 \cdot a^{0.3422} \cdot z^{0.7455} \tag{3}$$

We tested these two models on the three available datasets. Our experiments are summarized in Table 2, where the reference results of φ^*_{SG} (Eq. 1) are also recalled. Overall, φ^*_{SG} and the two generated functions φ^*_{G} and φ^*_{D} yield similar results in accuracy.

[2] http://creativemachines.cornell.edu/eureqa

Table 2 Accuracy of the approximation (r^2) of the φ-index by φ^*_{SG}, φ^*_{G} and φ^*_{D} with respect to three datasets

Dataset	Models for φ^*		
	φ^*_{SG} (Eq. 1) Reference	φ^*_{G} (Eq. 2) Learned on `EB_IS_2009`	φ^*_{D} (Eq. 3) Learned on `DBLP_2012`
Hevesy awardees (Schubert 2012a)	0.8484	0.8284 (−2.4 %)	0.8340 (−1.7 %)
`EB_IS_2009` (Cabanac 2012)	0.9211	0.9392 (+2.0 %)	0.9283 (+0.8 %)
`DBLP_2012`	0.8695	0.8472 (−2.6 %)	0.8699 (+0.0 %)

This suggests that machine learning based on symbolic regression failed to find a better model than the Schubert–Glänzel's φ^*_{SG} model.

In a second experiment, we used symbolic regression to learn the exponents of a and b in φ^*_{SG} (Eq. 1), thus letting $c = 1$. The φ^*_{C} model learned (Eq. 4) has $r^2 = 0.8405$ on the `DBLP_2012` dataset, which is lower than for φ^*_{SG} ($r^2 = 0.8695$). This suggests that the Schubert–Glänzel's φ^*_{SG} is more accurate than the two-exponent model discovered through symbolic regression.

$$\varphi^*_C = a^{0.2276} \cdot z^{0.6690} \tag{4}$$

Finally, we wondered whether another "embarrassingly simple relation"—*dixit* Schubert (2012a, p. 304)—than φ^*_{SG} could be found between an author's partnership ability and his/her number of co-authors (a) plus citation rate (z). Among the several hundred models that symbolic regression learned, we selected four solutions and discuss their complexity and accuracy on `DBLP_2012` with respect to the reference accuracy of φ^*_{SG} ($r^2 = 0.8695$).

The $\varphi^*_{SR_1}$ model (Eq. 5) is the simplest one regarding its complexity (i.e., type and number of operators). With $r^2 = 0.8461$, it is however 2.7 % less accurate than the reference accuracy. Refinements of this model through genetic algorithms led to a second model: $\varphi^*_{SR_2}$ (Eq. 6) achieves $r^2 = 0.9100$, which is 4.7 % better than the reference accuracy. Notice that such a gain in accuracy implied a much more complex formula. This is also the case of the $\varphi^*_{SR_3}$ model (Eq. 7) achieving a better $r^2 = 0.9136$, which is 5.1 % better than the reference. Likewise, the $\varphi^*_{SR_4}$ model (Eq. 8) yields $r^2 = 0.9227$, which is 6.1 % better than the reference.

$$\varphi^*_{SR_1} = \min\left(a^{\frac{1}{2}}, z^2\right) \tag{5}$$

$$\varphi^*_{SR_2} = \min\left((a \cdot \lfloor z \rfloor)^{0.359}, \left(z^{\tanh(z)}\right)^{\ln(a)}\right) \tag{6}$$

$$\varphi^*_{SR_3} = \min\left(\min(a,z) \cdot \sqrt{\lfloor a \rfloor^{0.415}}, \min\left((a \cdot z)^{0.357}, z^{a^{0.415}}\right)\right) \tag{7}$$

$$\varphi^*_{SR_4} = \min\left(a, 0.9455 + a \cdot z \cdot \operatorname{atan}\left(\frac{\sqrt{2.032 \cdot a}}{a + a \cdot z}\right) - a \cdot \operatorname{atan}\left(\frac{\sqrt{1.851 \cdot a}}{a + a \cdot z}\right)\right). \tag{8}$$

These machine learning experiments show that models learned through symbolic regression outperform the reference Schubert–Glänzel's φ^*_{SG} by a 6.1 % margin *only*. This gain comes with an extra cost in terms of formula complexity and lack of mathematical grounding. Indeed, although φ^*_{SG} is related to Paretian distributions (Glänzel 2006), the $\varphi^*_{SR_i}$ variants only result from natural selection applied to random formulas. These points suggest that the Schubert–Glänzel's φ^*_{SG} model is accurate and robust on the domain-wide bibliographic dataset of computer science.

Conclusion

Schubert (2012a) introduced the Hirsch-type φ-index to assess the partnership ability of authors. On a sample of 34 leading researchers awarded with the Hevesy medal, he also showed the consistency ($r^2 = 0.8484$) of the Schubert–Glänzel's φ^*_{SG} model of h-index with the empirical values of φ. Similar conclusions ($r^2 = 0.8845$) were reported in a study about 58 jazz performers (Schubert 2012b).

This article upscaled Schubert's (2012a) experiments with a dataset of a million computer scientists. Our results suggests that φ^*_{SG} is also consistent ($r^2 = 0.8695$) with φ on this larger bibliographic dataset of varied researcher profiles. Moreover, symbolic regression run on this million-author dataset discovered models with a gain in accuracy of 6.1 % at most ($r^2 = 0.9227$). Unlike φ^*_{SG}, these models do not rely on mathematical foundations though. Consequently, the Schubert–Glänzel's model φ^*_{SG} of the partnership ability φ-index appears to be superior regarding both its mathematical grounding and accuracy.

Appendix: SQL code developed to compute φ and φ^*_{SG}

We processed the bibliographic records using SQL (structured query language) with the Oracle relational database management system. The reader interested in data processing with SQL applied to scientometrics is referred to (Wolfram 2006; Mallig 2010).

Listing 1 Oracle SQL code used to compute φ and φ^*_{SG}

```
-- authorship = [idAuthor, idPaper]
create table authorship
(
  idAuthor  number,
  idPaper   number,
  constraint pk_authorship primary key(idAuthor, idPaper)
) ;

-- collaborations = [idAuthor1, idAuthor2, nbCoauthoredPapers, partRank]
create view collaborations as
  select a1.idAuthor as idAuthor1,
         a2.idAuthor as idAuthor2,
         count(*) as nbCoauthoredPapers,
         rank() over (partition by a1.idAuthor
                      order by count(*) desc, a2.idAuthor) partRank
  from authorship a1, authorship a2
  where a1.idPaper = a2.idPaper
    and a1.idAuthor <> a2.idAuthor
  group by a1.idAuthor, a2.idAuthor ;

-- phi = [idAuthor, phi, phiSG]
create view phi as
    select idAuthor1 as idAuthor,
           sum(case when nbCoauthoredPapers >= partRank then 1 else 0 end) as phi,
           -- a: total nbr of coauthors, z: mean nbr of occurrence of the coauthors
           1 * power(count(*), 1/3) * power(avg(nbCoauthoredPapers), 2/3) as phiSG
    from collaborations
    group by idAuthor1
  union all
    -- authors who never collaborated with anyone else
    select distinct idAuthor, 0, 0
    from authorship
    where idAuthor not in (select idAuthor1 from collaborations) ;
```

In Listing 1, we first create the `authorship` table to store the author-paper pairs. Then, the `collaborations` view computes the list of co-authors of each author, with the number of joint papers and associated partnership rank, as in Table 1. Finally, the `phi` view computes the φ and φ^*_{SG} values for each author, including those who never collaborated (hence $\varphi = 0$ and $\varphi^*_{SG} = 0$).

In Listing 2 demonstrates the computation of φ and φ^*_{SG} for Albert Einstein according to his collaborations listed in Table 1. First, author-paper pairs are inserted in the `authorship` table. Then, a `select` statement retrieves data from the `phi` view.

Listing 2 Computing of φ and φ^*_{SG} for Albert Einstein (Table 1)

```
-- 8 papers by Einstein (1) and Mayer (2)
insert into authorship values(1, 1021) ;
insert into authorship values(2, 1021) ;
-- ... continued ...
insert into authorship values(1, 1028) ;
insert into authorship values(2, 1028) ;

-- ... continued ...

-- 1 paper by Einstein (1) and Tolman (25)
insert into authorship values( 1, 1251) ;
insert into authorship values(25, 1251) ;

-- Values of phi for Albert Einstein
select *
from phi
where idAuthor = 1 ;

-- IDAUTHOR PHI PHISG
-- -------- --- -----
--        1   3 4.515038847927441173292130029260530034 58
```

References

Alonso, S., Cabrerizo, F., Herrera-Viedma, E., Herrera, F. (2009). *h*-Index: a review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics, 3*(4), 273–289. doi:10.1016/j.joi.2009.04.001.

Bar-Ilan, J. (2008). Informetrics at the beginning of the 21st century: a review. *Journal of Informetrics, 2*(1), 1–52. doi:10.1016/j.joi.2007.11.001.

Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics, 87*(3), 597–620. doi:10.1007/s11192-011-0358-1.

Cabanac, G. (2012). Shaping the landscape of research in information systems from the perspective of editorial boards: a scientometric study of 77 leading journals. *Journal of the American Society for Information Science and Technology, 63*(5), 977–996. doi:10.1002/asi.22609.

Chen, J., & Konstan, J. A. (2010). Conference paper selectivity and impact. *Communications of the ACM, 53*(6), 79–83. doi:10.1145/1743546.1743569.

Elmacioglu, E., & Lee, D. (2005). On six degrees of separation in DBLP-DB and more. *SIGMOD Record, 34*(2), 33–40. doi:10.1145/1083784.1083791.

Freyne, J., Coyle, L., Smyth, B., & Cunningham, P. (2010). Relative status of journal and conference publications in Computer Science. *Communications of the ACM, 53*(11), 124–132. doi:10.1145/1839676.1839701.

Glänzel, W. (2006). On the h-index: a mathematical approach to a new measure of publication activity and citation impact. *Scientometrics, 67*(2), 315–321. doi:10.1007/s11192-006-0102-4.

Hirsch, JE. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA, 102*(46), 16569–16572. doi:10.1073/pnas.0507655102.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. Cambridge: MIT Press.

Ley, M. (2002). The DBLP computer science bibliography: evolution, research issues, perspectives. In: A. H. F. Laender, A. L. Oliveira (eds.) *SPIRE'02 : Proceedings of the 9th international conference on string processing and information retrieval* (vol. 2476, pp. 1–10). Springer, LNCS. doi:10.1007/3-540-45735-6_1.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–324.

Mallig, N. (2010). A relational database for bibliometric analysis. *Journal of Informetrics, 4*(4), 564–580. doi:10.1016/j.joi.2010.06.007.

Rousseau, R. (2012). Comments on "A Hirsch-type index of co-author partnership ability". *Scientometrics, 91*(1), 309–310. doi:10.1007/s11192-011-0606-4.

Schmidt, M., Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81–85. doi:10.1126/science.1165893.

Schreiber, M., Malesios, C., Psarakis, S. (2012). Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators. *Journal of Informetrics, 6*(3), 347–358. doi:10.1016/j.joi.2012.02.001.

Schubert, A. (2012a). A Hirsch-type index of co-author partnership ability. *Scientometrics, 91*(1), 303–308. doi:10.1007/s11192-011-0559-7.

Schubert, A. (2012b). Jazz discometrics: a network approach. *Journal of Informetrics, 6*(4), 480–484. doi:10.1016/j.joi.2012.04.004.

Schubert, A., Glänzel, W. (2007) A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics, 1*(3), 179–184. doi:10.1016/j.joi.2006.12.002.

Wolfram, D. (2006). Applications of SQL for informetric frequency distribution processing. *Scientometrics, 67*(2), 301–313. doi:10.1007/s11192-006-0101-5.

Zhao, S. X., Rousseau, R., Ye, F. Y. (2011). h-Degree as a basic measure in weighted networks. *Journal of Informetrics, 5*(4), 668–677. doi:10.1016/j.joi.2011.06.005.